

Investigating Feature Characteristics of Pseudo-Speaker Data for Speaker Verification

© Hengyi Zou, Sayaka Shiota (Tokyo Metropolitan University)

1 Introduction

Deep neural embeddings such as ECAPA-TDNN [1] and RawNet3 [2] have advanced automatic speaker verification (ASV), yet their performance depends heavily on the diversity and size of training data. When training data is limited, augmentation methods such as noise addition and VTLP-based pseudo-speaker generation are essential [3]. However, the characteristics of pseudo-speakers generated by VTLP have not been thoroughly investigated. In this paper, we investigate the distribution and separability of VTLP-based pseudo-speakers to evaluate their effectiveness and limitations in improving ASV.

2 Methodology

2.1 Pseudo-speaker generation

In [3], it has been reported that the use of VTLP for pseudo-speaker generation improves ASV performance. VTLP can be represented as:

$$\omega' = \omega + 2\arctan \frac{\alpha \sin(\omega)}{1 - \alpha \cos(\omega)}. \quad (1)$$

In this formula, ω and α denote normalized frequency and expansion coefficient, respectively. In our previous work [3], the conventional VTLP-based method, applied a fixed warp factor (e.g., ± 0.1) across all speakers. While this method improves data diversity, it ignores individual acoustic characteristics and may limit the variability of augmented data. In the conventional approach, cosine similarity between original and VTLP-processed embeddings was used to assess speaker-level variation. Samples with insufficient divergence were either discarded or regenerated using adjusted warp factors until a similarity threshold was exceeded or a preset limit (α between $[-0.16, -0.1]$ and $[0.1, 0.16]$, with a step size of 0.01) was reached (see Fig. 1). This distance-aware adjustment yields high-variability pseudo-speakers while reducing redundancy and label confusion, enhancing both diversity and realism in the augmented dataset. These

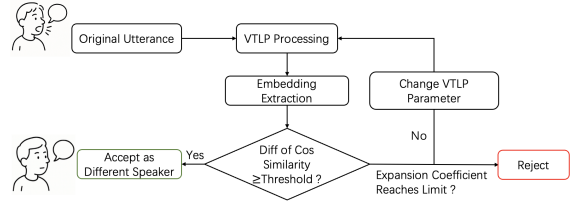


Fig. 1 Proposed pseudo-speaker generation process

pseudo-speaker utterances are then incorporated into the training set to improve the robustness of the speaker embedding model. By introducing high-variability but realistic samples, the system learns to better distinguish between speakers, which leads to improved ASV performance.

2.2 Embedding extraction and analysis

In recent ASV systems, speaker embeddings are typically used to represent utterances. In our analysis, embeddings are extracted and compared between original and augmented data. For each speaker, we compute the cosine similarity between original and VTLP-augmented embeddings and quantify the shift. Augmented samples exceeding a predefined similarity difference threshold (e.g., 0.35) are retained to ensure sufficient distinctiveness. The threshold was empirically determined based on preliminary experiments.

3 Experimental Condition

To investigate the speaker characteristics of pseudo-speakers, we trained a RawNet3-based [2] embedding model using a subset of the VoxCeleb1 dataset (400 speakers, 49,009 utterances). We employed the VoxCeleb.Trainer[4] implementation, using the default settings except for the number of training epochs, which was set to 400. Performance is measured using the equal error rate (EER) and the minimum detection cost function (MinDCF).

The following augmentation strategies were compared:

- Original: No augmentation applied.

Table 1 Number of utterances and speakers per condition

Conditions	Orig.	All	Select	Prop.
#utterances	49,009	147,027	73,555	113,695
#spkrs	400	1,200	1,143	1,191

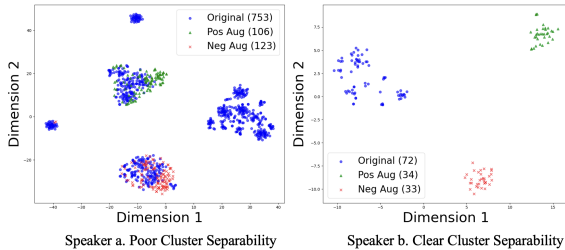


Fig. 2 t-SNE visualization of speaker embeddings.

- Noise: Add noise and reverberation to the original data using the MUSAN corpus and simulated Room Impulse Responses (RIRs).
- VTLP(All): Apply VTLP to all utterances.
- VTLP(Select): Retains pseudo-speakers whose cosine similarity exceeds a threshold.
- VTLP(Proposed): Re-generate high-variability pseudo-speakers by adjusting new VTLP parameters.

Table 1 shows the number of utterances and speakers under each condition. Compared to Select, the Proposed method achieves a substantial data increase by regenerating effective augmented samples.

4 Results and Discussion

To examine embedding-level characteristics, we visualized speaker embeddings via t-SNE. Figure 2 compares two cases: In (a), pseudo-speaker embeddings (green/red) substantially overlap with original speakers (blue), indicating limited inter-class variability. In (b), the proposed method yields more distinct clusters, suggesting improved separability and diversity.

We also evaluated each augmentation strategy using the RawNet3 model. Table 2 summarizes the EER and MinDCF under conditions with and without noise. The baseline (original data) achieves an EER of 2.73%. VTLP(All) significantly degrades performance (4.55%), while VTLP(Select) and VTLP(Proposed) improve it, with the proposed method achieving the best result (2.56%). However,

Table 2 EER / MinDCF under different augmentation conditions (with/without noise)

Conditions	EER	minDCF
Original	2.7307	0.132
Noise	2.1307	0.100
VTLP(All)	4.5511	0.219
VTLP(Select)	2.9168	0.149
VTLP(Proposed)	2.5652	0.123
VTLP(Proposed+Noise)	2.2549	0.110

with noise, its EER slightly worsens to 2.25%, compared to 2.13% for the original. This suggests that combining pseudo-speaker augmentation with noise may introduce excessive variability and hinder generalization. This may be attributed to the fact that VTLP introduces excessive spectral distortion, especially when applied to already noisy samples, leading to unrealistic variations in the embedding space.

5 Conclusion

We proposed a VTLP-based pseudo-speaker augmentation method that enhances speaker variability while preserving embedding separability. Cosine similarity-based selection and regeneration improve distinctiveness by filtering low-variability samples. Experiments on VoxCeleb1 demonstrate lower EER, MinDCF, and improved training stability compared to standard VTLP. And t-SNE visualizations show better cluster separability. Future work includes refining selection criteria and tuning thresholds.

6 Acknowledgment

This study was partially supported by JSPS KAKENHI (Grant Number JP24K14993), SCAT, and ROIS DS-JOINT (026RP2025), with funding provided to Prof. Sayaka Shiota.

References

- [1] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnns based speaker verification,” in *Proc. Interspeech*, pp. 3830–3834, 2020.
- [2] J. Jung, S. Kim, H. Shim, J. Kim, and H. Yu, “Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms,” in *Proc. Interspeech*, 2020.
- [3] H. Zou and S. Shiota, “Vocal tract length perturbation-based pseudo-speaker augmentation considering speaker variability for speaker verification,” in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–6, 2024.
- [4] ClovaAI, “Voxceleb_trainer.” https://github.com/clovaai/voxceleb_trainer, 2021. Accessed: 2025-07-11.